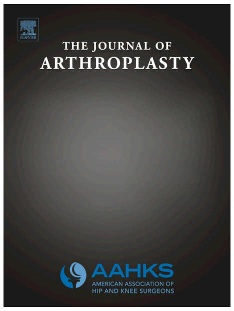


Deep Learning Artificial Intelligence Model for Assessment of Hip Dislocation Risk Following Primary Total Hip Arthroplasty from Postoperative Radiographs

Pouria Rouzrokh, MD, MPH, MHPE, Taghi Ramazanian, MD, Cody C. Wyles, MD, Kenneth A. Philbrick, PhD, Jason C. Cai, MBBS, Michael J. Taunton, MD, Hilal Maradit Kremers, MD, David G. Lewallen, MD, Bradley J. Erickson, MD, PhD

PII: S0883-5403(21)00167-4

DOI: https://doi.org/10.1016/j.arth.2021.02.028

Reference: YARTH 58671

To appear in: *The Journal of Arthroplasty*

Received Date: 30 September 2020

Revised Date: 4 February 2021

Accepted Date: 8 February 2021

Please cite this article as: Rouzrokh P, Ramazanian T, Wyles CC, Philbrick KA, Cai JC, Taunton MJ, Kremers HM, Lewallen DG, Erickson BJ, Deep Learning Artificial Intelligence Model for Assessment of Hip Dislocation Risk Following Primary Total Hip Arthroplasty from Postoperative Radiographs, *The Journal of Arthroplasty* (2021), doi: https://doi.org/10.1016/j.arth.2021.02.028.

Deep Learning Artificial Intelligence Model for Assessment of Hip Dislocation Risk Following Primary Total Hip Arthroplasty from Postoperative Radiographs

Running Title: CNN to predict dislocation in THA

Pouria Rouzrokh, MD, MPH, MHPE[1], Taghi Ramazanian, MD[2, 3], Cody C. Wyles, MD[2, 3], Kenneth A. Philbrick, PhD[1], Jason C. Cai, MBBS[1], Michael J. Taunton, MD[2, 3], Hilal Maradit Kremers, MD[2, 3, *], David G. Lewallen, MD[3], Bradley J. Erickson, MD, PhD[1]

[1]Department of Radiology, Radiology Informatics Laboratory, Mayo Clinic, 200 First Street SW, Rochester, MN, 55905, USA

[2]Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN, 55905, USA

[3]Department of, Orthopedic Surgery, Mayo Clinic, 200 First Street SW, Rochester, MN, 55905, USA

*** Please address all correspondence to**:

Hilal Maradit Kremers, M.D.
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
maradit@mayo.edu

Funding: This work was supported by the Mayo Foundation Presidential Fund, and the National Institutes of Health (NIH) [grant numbers R01AR73147 and P30AR76312]. The funding source had no role in the study design; collection, analysis and interpretation of the data; the writing of the report; or the decision to submit the article for publication.

Conflict of Interest: Dr. Wyles is an AAHKS Research Committee Member. Dr. Lewallen reports royalties and a paid consultant with Zimmer, Biomet; stock with Acuitive Technologies, Ketai Medical Devices; research support from Corin; board member of American Joint Replacement Registry, Orthopaedic Research and Education Foundation. Dr. Taunton is a paid consultant, and has royalties with DJO Global and from the publisher of the Journal of Bone and Joint Surgery. He is also on the governing board for the Journal of Arthroplasty and AAHKS.

Deep Learning Artificial Intelligence Model for Assessment of Hip Dislocation Risk Following Primary Total Hip Arthroplasty from Postoperative Radiographs

Running Title: CNN to predict dislocation in THA

Abstract

Background: Dislocation is a common complication following total hip arthroplasty (THA), and accounts for a high percentage of subsequent revisions. The purpose of this study was to illustrate the potential of a convolutional neural network (CNN) model to assess the risk of hip dislocation based on postoperative anteroposterior (AP) pelvis radiographs.

Methods: We retrospectively evaluated radiographs for a cohort of 13,970 primary THAs with 374 dislocations over 5 years of follow-up. Overall, 1,490 radiographs from dislocated and 91,094 from non-dislocated THAs were included in the analysis. A CNN object detection model (YOLO-V3) was trained to crop the images by centering on the femoral head. A ResNet18 classifier was trained to predict subsequent hip dislocation from the cropped imaging. The ResNet18 classifier was initialized with ImageNet weights and trained using FastAI (V1.0) running on PyTorch. The training was run for 15 epochs using ten-fold cross validation, data oversampling and augmentation.

Results: The hip dislocation classifier achieved the following mean performance: accuracy= 49.5(±4.1)%, sensitivity= 89.0(±2.2)%, specificity= 48.8(±4.2)%, positive predictive value= 3.3(±0.3)%, negative predictive value= 99.5(±0.1)%, and area under the receiver operating characteristic curve= 76.7(±3.6)%. Saliency maps demonstrated that the model placed the greatest emphasis on the femoral head and acetabular component.

Conclusions: Existing prediction methods fail to identify patients at high risk of dislocation following THA. Our radiographic classifier model has high sensitivity and negative predictive value, and can be combined with clinical risk factor information for rapid assessment of risk for dislocation following THA. The model further suggests radiographic locations which may be important in understanding the etiology of prosthesis dislocation. Importantly, our model is an illustration of the potential of automated imaging AI models in orthopedics.

Level of Evidence: Level III

Keywords: total hip arthroplasty; total hip replacement; dislocation; artificial intelligence; deep learning; convolutional neural network

Introduction

Dislocation is the most common early complication following primary THA and is one of the main indications for revision surgery[1, 2]. Based on pooled analysis of 4,633,935 primary THAs, the incidence of dislocation is estimated at 2.10% over an average follow-up of six years[3]. Dislocation is accompanied by severe pain, loss of limb function, need for revision surgeries, and an increase in treatment costs of up to 300%, compared to an uncomplicated THA[4].

Identifying patients at risk of dislocation following primary THA is important for surgical planning, postoperative restrictions, and rehabilitation protocols which may reduce the risk of postoperative hip dislocation. Several risk factors are associated with increased risk of dislocation. At the patient level, age higher than 70 years, body mass index (BMI) greater than 30 kg/m2, comorbidities like neuromuscular disorders or cognitive impairment, and previous surgeries including spinal fusion are associated with an elevated dislocation risk[3, 5-7]. Among surgery-related factors, component malpositioning during surgery and a posterior surgical approach, especially without anatomical repair of the posterior capsule and the external rotators, increase the risk of dislocation. Several studies have suggested that the positioning of the acetabular component affects dislocation risk. Dislocation is more common with smaller femoral head diameters, whereas implant features such as dual mobility acetabular designs are associated with reduced risk of dislocation[8-11].

Plain radiographs are used to evaluate for post-operative complications such as malposition loosening, and periprosthetic fracture[12, 13]. Previous studies have used post-operative anteroposterior (AP) pelvis radiographs to measure femoral and acetabular offset and/or, to measure inclination and anteversion angles to determine acetabular component position[14]. Others have investigated the dislocation risk based on measuring hip adduction and pelvic obliquity deformity on pre-operative pelvis radiographs[15].

Convolutional neural networks (CNNs) are the current state-of-the-art artificial intelligence (AI) techniques for fully automated medical image analysis[16]. These networks “learn” to predict outcomes or measures by looking for low-level image features such as edges and curves and then building up to more abstract concepts through a series of convolutional layers[17]. Although researchers can train CNNs on medical datasets from scratch, this approach is usually hindered by the limited number of available images. CNNs generally require large datasets to achieve high-level performance, but “transfer learning” can help to overcome this barrier[18]. In transfer learning, CNNs initially learn to identify predictive imaging features by being trained on a large dataset. Subsequently they are further trained on a smaller dataset to learn to map the learned features predict transferred task.

Recent studies have used non-imaging AI models to predict dislocations following THA[19], whereas imaging AI models have been used to detect other THA complications[20]. To our knowledge, no study has yet reported the application of an imaging AI model to assess the risk of dislocations following THA. In this study, we introduce a CNN model to classify patients based on their risk for dislocation using postoperative anteroposterior (AP) pelvis radiographs. Although in practice, surgical decisions are not made by solely relying on imaging data, we design our study to illustrate the potential of imaging AI models to predict hip dislocation as a rare and multi-factorial outcome.

Materials and Methods

Assembling the Imaging Dataset

Following Institutional Review Board (IRB) approval, we retrospectively assembled a cohort of 13,970 primary THAs performed between 2000 and 2017 at a single academic institution. Indications for THA were osteoarthritis, rheumatoid arthritis, or avascular necrosis. Over a mean 5.0 years of follow-up, 374 (2.7%) sustained a dislocation compared to 13,596 (97.3%) who did not dislocate during follow-up (hereafter called: normal). Females constituted 62.5% of the dislocation class and 51.0% of the normal class. This difference was statistically significant (P-value: <0.001).

Figure 1 summarizes the methodology of the study. 97,934 AP pelvis radiographs were retrieved for the study population, taken at least one day after the surgery date and at least one day before the possible dislocation date. An orthopedic surgeon reviewed the images to ensure that no dislocation had been present at the time of imaging. A total of 5,350 images were excluded due to artifacts, poor visibility of implants or bones, or abnormal cropping. Overall, 1,490 AP radiographs from the dislocation class and 91,094 from the normal class were included for our study. Table 1 compares descriptive variables between classes in our imaging dataset. Within this dataset, age, weight, and height of patients at the time of surgery were slightly different between classes, and the frequency of females in dislocation class was significantly higher than the normal class (as was in our study population).

Prediction of and Cropping the Region of Interest

A CNN object detection model (YOLO-V3) was trained to crop the images by centering on the femoral head and help the dislocation AI model focus on the most relevant parts of the image. YOLO-V3 can be trained to detect bounding boxes of interest within images[21]. Before training YOLO-V3, we first zero-padded (added pixels with value of zero) all images to make them square-shaped and then resized them to 512×512 pixels. For annotation, we manually determined the bounding boxes on 10,000 AP pelvis radiographs from left and right sided THAs. Annotations were done in a way that the medial border of the box was in line with body midline (through the pubic symphysis), the lateral border was tangent to the greater trochanter, the inferior border was tangent to the inferior pubic ramus, and the superior border was tangent to the superior part of the acetabular hardware (Figure 2a). Training, validation, and test subsets included 8500, 1000, and 500 images, respectively. YOLO-V3 was then was trained for 15 epochs, with a batch size of 4 and a learning rate of 0.0001 with pooling weights from a pre-trained model on the Microsoft Common Objects in Context (COCO) dataset[22]. Training was done on an NVIDIA GeForce 1080Ti GPU with 11 Gigabytes of RAM using the ImageAI library (V1.0) running on Tensorflow. To apply YOLO-V3, the predicted bounding boxes for images were dilated by 10% on the superior, medial, and lateral sides before final cropping (if present in original radiograph) to ensure that a broader view of the pubic symphysis, acetabular component and femur trochanters was included (Figure 2b). The cropped images were again zero-padded to a square shape and resized to 224 × 224 pixels (Figure 2c). Final images were also normalized with respect to their individual mean and standard deviation.

Assessment of hip dislocation

Model, Initialization and Training

We created a ResNet18 model with initial weights pooled from a model pre-trained on the ImageNet database. We trained the network’s output layer for 15 epochs, with a batch size of 16, a learning rate of 0.0001 and using the Adam optimizer. All layers of the model were then fine-tuned for five more epochs using a learning rate slice adjusted based on the FastAI Learning Rate Finder tool. We used binary cross entropy as our loss function and weighted it 25 times more for the dislocation class than the normal class. During training, the model with the highest Area Under the Curve (AUC) and sensitivity of at least 80% in detecting dislocation class on the validation data was saved as the final model. All above numeric choices (also called hyperparameters) were chosen based on best knowledge of deep learning literature and our experimental trainings. We trained our ResNet18 model on an NVIDIA Tesla V-100 GPU with 32 Gigabytes of RAM using FastAI (V1.0) running on PyTorch.

Ten-fold Cross-validation

Performance of the ResNet18 model was assessed using ten-fold cross-validation. Data was split between folds with stratified randomization based on the data classes. Within each fold, the training, validation, and testing subset split was 90%, 5%, and 5%, respectively. Every image was present in the training subset for nine folds and belonged to either validation or test subsets for one fold. Images were split by Patient IDs, so that no images in different subsets belonged to the same patient. As the number of available images in our imaging dataset varied between patients, the number of unique images allocated to subsets was not the same among folds. An average fold included 83,331(±42) images (Dislocated=1,341 (1.6%), Normal=81,989(98.4%)).

In addition, images from the dislocation class were over-sampled (copied) to match the number of the normal class in training and validation subsets of all folds. The average oversampling factor was 61 and 75 times for the training and validation subsets, respectively. To make over-sampling more effective, copied images were also slightly changed (augmented) compared to the original images. Augmentation included one or more of: horizontal flipping, rotation between ± 10°, and a maximum of 10% zooming for each image. The test subset in all folds remained imbalanced to represent the real-world data.

Among the training and validation subsets, images from female patients were 1.25(±0.03) and 1.54(±0.58) times more frequent in the dislocation class than the normal class, respectively. To ensure that patient gender did not affect the reported statistics of the model, the random allocation of images to the test subset was stratified by gender so that the male/female ratio was the same for the normal and dislocation classes.

Outputs and Statistics

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of each fold's model was measured on the test subset for that fold. By doing so, we assumed happening of dislocation as our positive outcome. We also reported the mean and standard deviation of the above statistics across all folds. A paired sample t-test was used to test if the results in that fold differed from the overall average result. The confusion matrix, training loss, validation loss and receiver operating characteristic (ROC) curves were plotted for all folds. For each fold, we also applied the model on two distinct test samples that included exclusively male or exclusively female patients. Both samples had a dislocation frequency of 2% and included no images from the training subset. This helped to compare our model's performance when applied to images from different genders.

We created saliency maps for one representative image from each fold's test subset to demonstrate that our model is making decisions based on meaningful features within the images, and its performance is therefore reliable.

Independent t-test and Pearson chi-square tests were done using the SciPy statistical package (V1.4.1) in python (V3.6), and p-value <0.05 was considered as significant.

Results

The YOLO-V3 model achieved a mean Average Precision (mAP) of 99.3% and 99.1% in detecting regions of interest for right and left pelvis, respectively. Overall, mAP for the model was 99.2%.

Table 2 summarizes the ten-fold performance of the model over the test subset. Due to class imbalance preserved in the test subset, 2% of images in the test subset were from the dislocation class, compared to 98% of images from the normal class. On average, the ResNet18 classifier achieved an accuracy of 49.5(±4.1)%, sensitivity of 89.0(±2.2)%, specificity of 48.8(±4.2)%, PPV of 3.3(±0.3)%, NPV of 99.5(±0.1)%, and AUC of 76.7(±3.6)% across all folds. Neither of the folds' results were different from the reported average (p-value > 0.6). Figure 3a shows the ten-fold average ROC curve for the ResNet18 classifier applied over the test subset. Loss curves and the confusion matrix for one fold (fold 2) is plotted in Figures 3b and 3c. Supplementary Figures 1-3 include ROC curves, confusion matrices, and loss curves for all the folds.

Table 3 compares the ten-fold average performance of the classifier model when applied on images from males and females separately. While the model was more sensitive in predicting dislocation among females, it was more specific when applied to images from the male patients. The NPV of the model was not different between two groups.

Figure 4 shows saliency maps for representative instances of correct classifications from the normal and dislocation classes. Colored regions on the saliency maps denote the relative influence of individual pixels on the model's decisions, where the red pixels highlight the most influential regions. Saliency maps provide evidence that the model placed considerable emphasis on the femoral head and acetabular component of implant, while the pelvic rami and the greater trochanter of the femur were also emphasized in some decisions.

Discussion

Dislocation is a rare complication following THA, and is associated with pain and reduced function, subsequent revision surgeries, and substantial healthcare costs. In this study, we trained a CNN to classify THA patients based on their risk for dislocation from single postoperative AP pelvis radiographs (without considering other patient data). Although surgeons can rely on other imaging modalities (like computed tomography) to investigate the risk of dislocation, we solely included X-rays as they are the routine imaging modality in THA and thus, are appropriate for screening purposes. Dislocation images constituted about 2% of total images in our dataset and we proportionally preserved this imbalance in our test subset. Although the average accuracy and specificity of our model is not high, it detected about 90% of patients who dislocated in the future based on a single postoperative radiograph. On the other hand, the chance of dislocation in patients who have been classified as "to-be-normal" by the model is approximately 0.5%, representing a four-fold decrease from the baseline dislocation rate. We acknowledge that better predictive performance can be achieved by accounting for various dislocation-related clinical risk factors. Nevertheless, our imaging AI model and more sensitive future extensions can be incorporated with demographic, clinical and surgical risk factors for rapid screening of patients at high risk for dislocation following THA. Considering the burden of follow-up visits and the reported drawbacks of hip precautions[23], patients classified as “to-be-normal” by the model may be treated with fewer restrictions.

Our model was more sensitive in detecting dislocation in female patients, and it was more specific in male patients. Such difference can be explained by the characteristics of our study population. In our study population and the imaging dataset, the male/female ratio is significantly lower in the dislocation class than in the normal class. This might have helped the model to be more expert in detecting patterns of dislocation in female patients. Also, the model possibly relies on different imaging features to predict dislocation when applied on different genders. The gender-related differences in sensitivity and specificity of our model should be considered when applying the model in the real clinical setting. Nevertheless, the model has high negative predictive value regardless of patient gender.

Saliency maps are tools that highlight the individual pixels’ importance in the decision making of the model[24, 25]. Although, deep learning techniques cannot currently clarify why CNN models make specific decisions, saliency maps can give us some clues to guess the reasoning behind models’ decisions. Saliency maps generated for representative images illustrate that our model relied on several anatomical zones in predicting dislocations. The most consistent zone of interest on saliency maps was the area around the femoral head and the acetabular component. The model may be detecting imaging features associated with orientation of the acetabular component, and the size of the femoral head relative to that component. Both of these factors are known to influence the risk of dislocation[3, 26]. The second intuitive zone was the greater trochanter and the area superior to it. The model is likely using this zone to learn intuitions about the femoral offset. The femoral offset can influence the tension on the abductor muscles and the propensity of the femur to impinge on the pelvis during extreme movements[14]. The last intuitive zone in saliency maps is the pelvic rami. The shape of the pelvic rami may indirectly imply the medialization of the acetabular component. Likewise, it may relate to the shape of the obturator foramen, which is dynamically influenced by the spinal flexibility. Change of spinal flexibility (e.g. after spinal fusion) may increase the risk for dislocation following THA[27, 28].

Our study had several challenges. First, the number of radiographs in the dislocation class was smaller than the normal class in our dataset (1,490 images vs. 91,094 images). As only < 3% of THAs had sustained a dislocation in our study, we regarded our dataset as highly imbalanced. Class imbalance can have detrimental effects on CNN training as the model may learn to do the easy task, i.e., to learn to classify all examples as the more frequent class and ignore the minority class[29]. Also, class imbalance makes our model more prone to overfitting. Overfitting implies a situation in the training when the model predictions fail to generalize to non-training data (Figure 3b and Supplementary Figure 3). To address class imbalance and overfitting in our training, we used various strategies including transfer learning from a model with fewer parameters (ResNet18, instead of the larger ResNet34, 50 or 101), ten-fold cross-validation, data over-sampling, and data augmentation. Also, we used a YOLO model to crop X-rays before feeding them to the final classifier model. If we had left X-rays uncropped, the dislocation AI model would have a more difficult task to find informative imaging features and could be hit more by overfitting. Second, the number of available images for different patients varied in our dataset. As we needed to separate our folds and datasets at the patient level, this limitation was a barrier to preserve an equal gender-ratio between subsets of different folds; otherwise, some folds would be significantly larger or smaller than others. This also prevented us from including other X-ray views in our study, as not all dislocated patients had available x-rays in all views. Third, we limited our dataset to patients with osteoarthritis, rheumatoid arthritis, and avascular necrosis as the underlying indications for THA. Patients with fractures or tumors were excluded, as hardware used in those patients could significantly differ from other THAs. Finally, we only relied on single AP pelvis radiographs to assess the risk of dislocations. Using serial or/and non-AP radiographs, standardizing radiographs based on factors like weight-bearing, and feeding non-imaging clinical data to the model will likely improve the classification performance.

Conclusion

Our study illustrates the potential of imaging AI models to investigate the risk of THA complications. We report an AI model that can do a meaningful classification of dislocations following THA based on single AP pelvis radiographs. It also introduces several zones of interest that may convey important etiological insights about dislocation outcomes. We are currently exploring the inclusion of demographic data, non-imaging clinical data, and surgical data to further improve the classification performance of the model. Finally, we invite other orthopedic sites to share their datasets and collaborate to build pooled datasets of THA imaging studies. Notably, deep learning models need sufficiently large datasets and data sharing is the most practical – if not the only – way to improve performance when dealing with rare outcomes like hip dislocation.

References

1. Salassa T, Hoeffel D, Mehle S, Tatman P, Gioe TJ. Efficacy of Revision Surgery for the Dislocating Total Hip Arthroplasty: Report From a Large Community Registry. Clinical Orthopaedics and Related Research®. 2014;472(3):962-7.
2. Bozic KJ, Kurtz SM, Lau E, Ong K, Vail TP, Berry DJ. The epidemiology of revision total hip arthroplasty in the United States. The Journal of bone and joint surgery American volume. 2009;91(1):128-33.
3. Kunutsor SK, Barrett MC, Beswick AD, Judge A, Blom AW, Wylde V, Whitehouse MR. Risk factors for dislocation after primary total hip replacement: a systematic review and meta-analysis of 125 studies involving approximately five million hip replacements. The Lancet Rheumatology. 2019;1(2):e111-e21.
4. Abdel MP, Cross MB, Yasen AT, Haddad FS. The functional and financial impact of isolated and recurrent dislocation after total hip arthroplasty. The bone & joint journal. 2015;97-B(8):1046-9.
5. Dargel J, Oppermann J, Brüggemann G-P, Eysel P. Dislocation following total hip replacement. Deutsches Arzteblatt international. 2014;111(51-52):884-90.
6. Soong M, Rubash HE, Macaulay W. Dislocation after total hip arthroplasty. The Journal of the American Academy of Orthopaedic Surgeons. 2004;12(5):314-21.
7. Brooks PJ. Dislocation following total hip replacement: causes and cures. The bone & joint journal. 2013;95-B(11 Suppl A):67-9.
8. Lewinnek GE, Lewis JL, Tarr R, Compere CL, Zimmerman JR. Dislocations after total hip-replacement arthroplasties. The Journal of bone and joint surgery American volume. 1978;60(2):217-20.
9. Reina N, Putman S, Desmarchelier R, Sari Ali E, Chiron P, Ollivier M, Jenny JY, Waast D, Mabit C, de Thomasson E, Schwartz C, Oger P, Gayet LE, Migaud H, Ramdane N, Fessy MH. Can a target zone safer than Lewinnek's safe zone be defined to prevent instability of total hip arthroplasties? Case-control study of 56 dislocated THA and 93 matched controls. Orthopaedics & traumatology, surgery & research : OTSR. 2017;103(5):657-61.
10. Pollard JA, Daum WJ, Uchida T. Can simple radiographs be predictive of total hip dislocation? The Journal of arthroplasty. 1995;10(6):800-4.

11. Esposito CI, Gladnick BP, Lee Y-Y, Lyman S, Wright TM, Mayman DJ, Padgett DE. Cup position alone does not predict risk of dislocation after hip arthroplasty. The Journal of arthroplasty. 2015;30(1):109-13.

12. Vanrusselt J, Vansevenant M, Vanderschueren G, Vanhoenacker F. Postoperative radiograph of the hip arthroplasty: what the radiologist should know. Insights into imaging. 2015;6(6):591-600. Epub 2015/10/20.

13. Mushtaq N, To K, Gooding C, Khan W. Radiological Imaging Evaluation of the Failing Total Hip Replacement. Frontiers in Surgery. 2019;6:35-.

14. Bhaskar D, Rajpura A, Board T. Current Concepts in Acetabular Positioning in Total Hip Arthroplasty. Indian journal of orthopaedics. 2017;51(4):386-96.

15. Liu Q, Cheng X, Yan D, Zhou Y. Plain radiography findings to predict dislocation after total hip arthroplasty. Journal of Orthopaedic Translation. 2019;18:1-6.

16. Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep Learning in Radiology: Does One Size Fit All? Journal of the American College of Radiology : JACR. 2018;15(3 Pt B):521-6.

17. Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. Procedia Computer Science. 2018;132:679-88.

18. Li X, Grandvalet Y, Davoine F, Cheng J, Cui Y, Zhang H, Belongie S, Tsai Y-H, Yang M-H. Transfer learning in computer vision tasks: Remember where you come from. Image and Vision Computing. 2020;93:103853-.

19. Alastruey-López D, Ezquerra L, Seral B, Pérez MA. Using artificial neural networks to predict impingement and dislocation in total hip arthroplasty. Computer methods in biomechanics and biomedical engineering. 2020:1-9.

20. Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. Detecting mechanical loosening of total hip replacement implant from plain radiograph using deep convolutional neural network. 2019.

21. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. 2018.

22. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P. Microsoft COCO: Common Objects in Context. 2014.

23. Barnsley L, Barnsley L, Page R. Are Hip Precautions Necessary Post Total Hip Arthroplasty? A Systematic Review. Geriatric orthopaedic surgery & rehabilitation. 2015;6(3):230-5.

24. Mundhenk TN, Chen BY, Friedland G. Efficient Saliency Maps for Explainable AI. 2019.

25. Philbrick KA, Yoshida K, Inoue D, Akkus Z, Kline TL, Weston AD, Korfiatis P, Takahashi N, Erickson BJ. What Does Deep Learning See? Insights From a Classifier Trained to Predict Contrast Enhancement Phase From CT Images. American Journal of Roentgenology. 2018;211(6):1184-93.

26. Biedermann R, Tonin A, Krismer M, Rachbauer F, Eibl G, Stöckl B. Reducing the risk of dislocation after total hip arthroplasty: the effect of orientation of the acetabular component. The Journal of bone and joint surgery British volume. 2005;87(6):762-9.

27. McKnight BM, Trasolini NA, Dorr LD. Spinopelvic Motion and Impingement in Total Hip Arthroplasty. The Journal of arthroplasty. 2019;34(7S):S53-S6.

28. Esposito CI, Carroll KM, Sculco PK, Padgett DE, Jerabek SA, Mayman DJ. Total Hip Arthroplasty Patients With Fixed Spinopelvic Alignment Are at Higher Risk of Hip Dislocation. The Journal of Arthroplasty. 2018;33(5):1449-54.

29. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis. 2002;6:429-49.

Figure Legends

Figure 1. Overview of the study.

Figure 2. [TO BE PRINTED IN COLOR]: Object detection YOLO-V3 model. (A) Zero-padding, resizing to 512 × 512 pixel size and initial annotation (B) Extension of detected area by ten percent towards the superior, medial and lateral sides (C) Final cropping, zero padding and resizing to 224 × 224 pixel size.

Figure 3. [TO BE PRINTED IN COLOR]: (a) ROC Curve showing the average performance of ResNet18 classifier model applied over the validation subsets in all folds (b) Training and validation loss curves for training the ResNet18 Classifier model in fold 2. The red dashed line represents the point where the best model was saved during the training (c) Confusion matrix showing the performance of ResNet18 classifier model applied over the test subset in fold 2.

Figure 4. [TO BE PRINTED IN COLOR]: Saliency maps for the trained models in each of the ten-folds (a) Example saliency maps for true predictions of the dislocation class (b) Example saliency maps for true predictions of the normal class.

Supplemental Figures

sFigure 1. Receiver operating characteristic curves showing the performance of ResNet18 classifier model applied over the validation subsets in all folds.

sFigure 2. Confusion matrices showing the performance of ResNet18 classifier model applied over the test subsets in all folds.

sFigure 3. Training and loss curves for training the ResNet18 Classifier model in all folds. The red dashed line implies the points where the best models were saved during the training. A lack of improvement in validation performance (beyond the red dashed line) indicates overfitting.

Table 1. Distribution of pelvis images from normal and dislocation classes. The last column shows the P-value of independent t-test or Pearson chi-square tests comparing the two classes.

Variable (Unit)	Normal Class	Dislocation Class	P-value
Total Hip Arthroplasty surgeries (Number (%) of Total)	10713 (97.58)	266 (2.42)	-
Unique Patients (Number (%) of Total)	8822 (97.20)	254 (2.80)	-
Pelvis Images (Number (%) of Total)	91094 (98.40)	1490 (1.60)	-
THA Side (Right/Left Ratio)	1.18	1.03	0.31
Gender (Male/Female Ratio)	0.96	0.45	< 0.001
Age at surgery (Median ± IQR of years)	68 (16)	66 (20)	0.008
Weight (Median ± IQR of kilograms)	85 (28)	83 (24)	0.002
Height (Median ± IQR of meters)	1.69 (0.15)	1.66 (0.13)	0.048
Follow-up duration (Median ± IQR of years)	5.01 (4.22)	4.94 (4.25)	0.783
Time to dislocation (Median ± IQR of years)	-	2.62 (3.51)	-

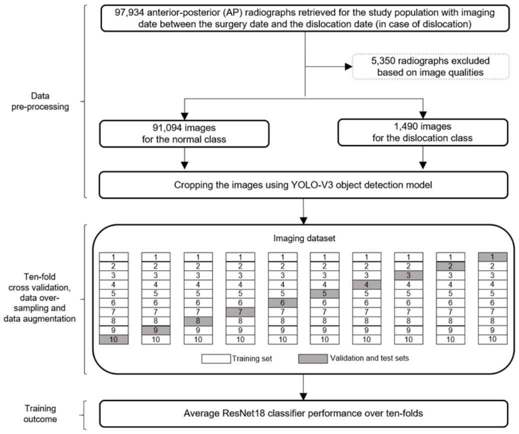
IQR, interquartile range

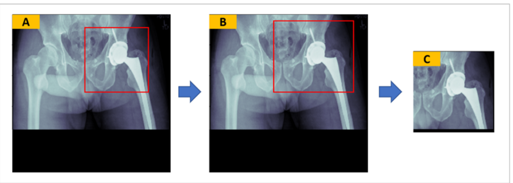
Table 2. Performance measures of ResNet18 classifier applied over ten-fold test subsets

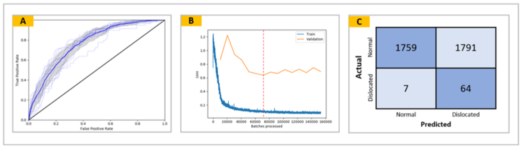
Fold	Accuracy (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	ROC Area Under the Curve (%)
1	45.65	87.50	44.91	2.74	99.51	72.30
2	50.35	90.14	49.55	3.45	99.60	79.10
3	49.50	87.18	48.74	3.29	99.48	74.50
4	46.68	93.15	45.75	3.32	99.70	81.70
5	46.18	91.03	45.28	3.22	99.61	79.90
6	52.61	86.11	51.94	3.46	99.47	77.70
7	54.98	88.00	54.32	3.71	99.56	78.50
8	50.48	91.55	49.66	3.51	99.66	77.90
9	56.97	85.92	56.39	3.79	99.50	77.00
`10	42.09	89.66	41.17	2.86	99.52	68.10
Mean	49.55	89.02	48.77	3.34	99.56	76.67
Standard deviation	4.11	2.22	4.20	0.30	0.07	3.63

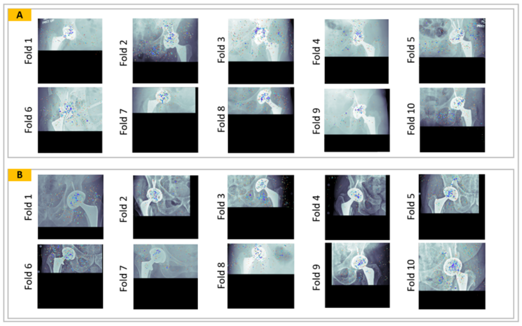
Table 3. Ten-fold mean (standard deviation) of the prediction model's performance when applied to image samples from exclusively male or exclusively female total hip arthroplasty patients. Samples had a dislocation frequency of 2% and included no images from the training subset.

Index	Male Patients	Female Patients	P-value
Sensitivity	84.9 (7.0)	90.0 (3.9)	0.044
Specificity	61.8 (5.2)	43.6 (5.6)	0.001
Positive Predictive Value	4.30 (0.4)	2.95 (0.4)	0.001
Negative Predictive Value	99.5 (0.2)	99.6 (0.2)	0.346
Accuracy	62.3 (5.0)	44.8 (5.5)	0.001


97,934 anterior-posterior (AP) radiographs retrieved for the study population with imaging date between the surgery date and the dislocation date (in case of dislocation)
5,350 radiographs excluded based on image qualities
Data pre-processing
91,094 images for the normal class
1,490 images for the dislocation class
Cropping the images using YOLO-V3 object detection model
Ten-fold cross validation, data over-sampling and data augmentation
Imaging dataset
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
Training set
Validation and test sets
Training outcome
Average ResNet18 classifier performance over ten-folds

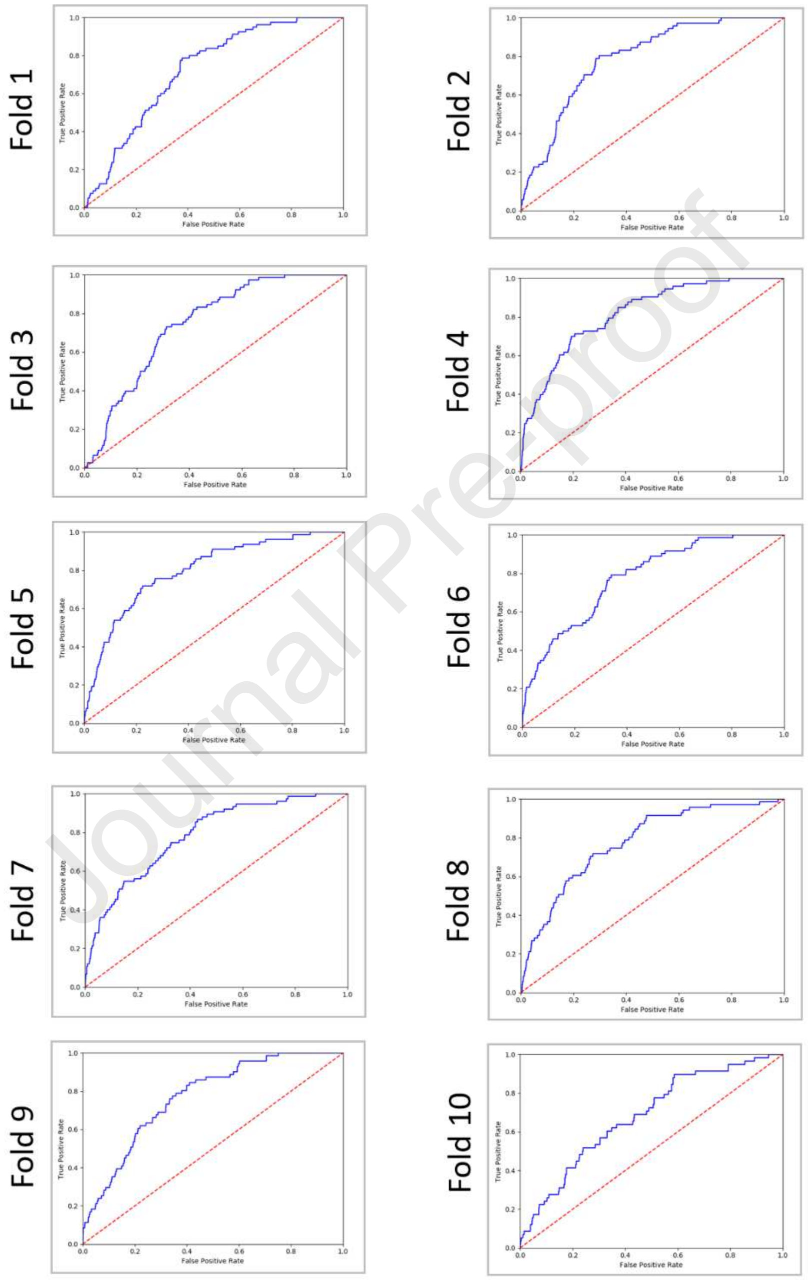

A
B
C


A
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
B
Train
Validation
Loss
Batches processed
0
20000
40000
60000
80000
100000
120000
140000
160000
0.2
0.4
0.6
0.8
1.0
1.2
C
Actual
Normal
Dislocated
1759
1791
7
64
Normal
Dislocated
Predicted

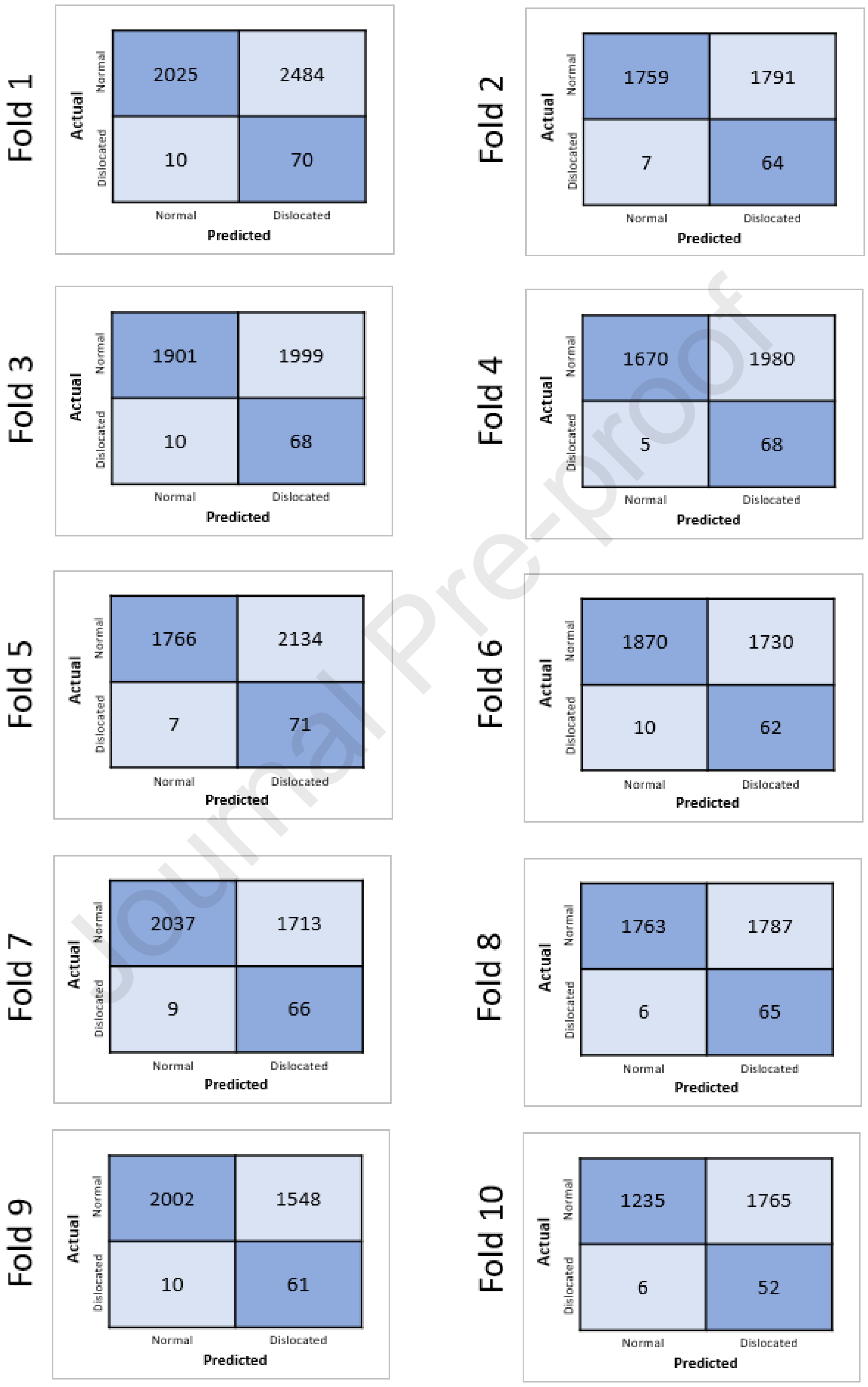

A
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10
B
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10

Supplemental Material

sFigure 1. Receiver operating characteristic curves showing the performance of ResNet18 classifier model applied over the validation subsets in all folds:



sFigure 2. Confusion matrices showing the performance of ResNet18 classifier model applied over the test subsets in all folds:



Fold 1

		Predicted	
		Normal	Dislocated
Actual	Normal	2025	2484
	Dislocated	10	70

Fold 2

		Predicted	
		Normal	Dislocated
Actual	Normal	1759	1791
	Dislocated	7	64

Fold 3

		Predicted	
		Normal	Dislocated
Actual	Normal	1901	1999
	Dislocated	10	68

Fold 4

		Predicted	
		Normal	Dislocated
Actual	Normal	1670	1980
	Dislocated	5	68

Fold 5

		Predicted	
		Normal	Dislocated
Actual	Normal	1766	2134
	Dislocated	7	71

Fold 6

		Predicted	
		Normal	Dislocated
Actual	Normal	1870	1730
	Dislocated	10	62

Fold 7

		Predicted	
		Normal	Dislocated
Actual	Normal	2037	1713
	Dislocated	9	66

Fold 8

		Predicted	
		Normal	Dislocated
Actual	Normal	1763	1787
	Dislocated	6	65

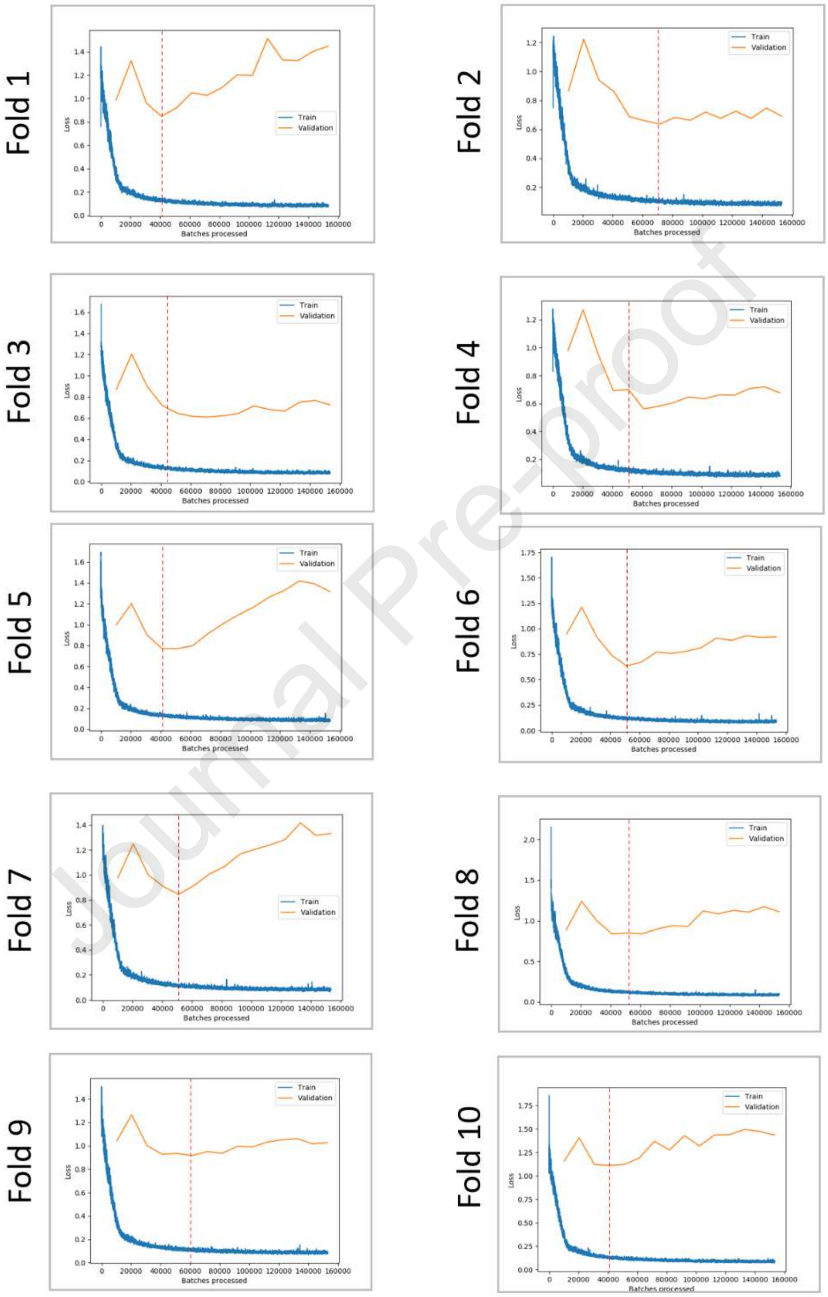
Fold 9

		Predicted	
		Normal	Dislocated
Actual	Normal	2002	1548
	Dislocated	10	61

Fold 10

		Predicted	
		Normal	Dislocated
Actual	Normal	1235	1765
	Dislocated	6	52

sFigure 3. Training and loss curves for training the ResNet18 Classifier model in all folds. The red dashed line implies the points where the best models were saved during the training. A lack of improvement in validation performance (beyond the red dashed line) indicates overfitting.


Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10
Loss
Batches processed
Train
Validation